Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI

Year: 2020 | Citations: 531 | Authors: Alon Jacovi, Ana Marasovi■, Tim Miller, Yoav Goldberg

---

Abstract

Trust is a central component of the interaction between people and AI, in that 'incorrect' levels of trust may cause misuse, abuse or disuse of the technology. But what, precisely, is the nature of trust in AI? What are the prerequisites and goals of the cognitive mechanism of trust, and how can we promote them, or assess whether they are being satisfied in a given interaction? This work aims to answer these questions. We discuss a model of trust inspired by, but not identical to, interpersonal trust (i.e., trust between people) as defined by sociologists. This model rests on two key properties: the vulnerability of the user; and the ability to anticipate the impact of the AI model's decisions. We incorporate a formalization of 'contractual trust', such that trust between a user and an AI model is trust that some implicit or explicit contract will hold, and a formalization of 'trustworthiness' (that detaches from the notion of trustworthiness in sociology), and with it concepts of 'warranted' and 'unwarranted' trust. We present the possible causes of warranted trust as intrinsic reasoning and extrinsic behavior, and discuss how to design trustworthy AI, how to evaluate whether trust has manifested, and whether it is warranted. Finally, we elucidate the connection between trust and XAI using our formalization.